Unit 06. Evaluation

Contents

1. Introduction

Evaluation

Evaluation involves reviewing, trying out or testing a design idea, a product or a service to discover whether it meets some criteria.

- These criteria will often be linked to the **guidelines for good design** and **usability** (effectiveness, efficiency, user satisfaction) or **user experience** (enjoyment, engagement and aesthetic appreciation) or specific characteristic (accessibility, etc.).

Evaluation and human-centered design

Hartson, Hix, 1989

Evaluation is central to human-centred design and is undertaken throughout the design process whenever a designer needs to check an idea, review a design concept or get reaction to a physical design (see star model).

Objectives of the unit

- Appreciation of the use of a range of generally applicable **evaluation techniques** used **with and without users**. In particular:
- Understanding and use **expert-based evaluation methods (without users)**.
- Understanding and use **user-based evaluation methods** .
- Understanding and use **data analytics methods**.

Introduction

- Evaluation of different types of systems and contexts, may offer particular challenges. For example, evaluating **mobile devices** or of **interaction with wearable devices**.
- How is evaluation related to the other key activities of UX design; **understanding**, **envisionment** and **design**?
  - Many of the **techniques used for understanding** are applicable to evaluation.
  - Evaluation is critically dependent on the **form of envisionment** used to represent the system.

Evaluation activities during design

In the human-centred approach to design, we evaluate designs right from the earliest idea. For example:

- Early ideas for a service can be discussed with other designers in a team meeting.
- **Mock-ups** can be quickly reviewed, and later in the design process, more **realistic prototyping** and testing of a partially finished system can be evaluated with users.
- Statistical evaluations of the **near-complete product** or service in its intended setting can be undertaken.
- Once the completed system is **fully implemented**, designers can evaluate alternative interface designs by gathering data about system performance (data analytics).

Three main types of evaluation

1. **expert-based** methods (with no user participation).
2. **participant-based** methods, also called 'user testing'.
3. **data analytics** methods on implemented systems.

- **Expert-based** methods will often pick up significant usability or UX issues quickly, but experts will sometimes miss detailed issues that real users find difficult.
- **Participant methods** must be used at some point in the development process to get real feedback from users.
- Both **expert-based and participant-based methods** can be conducted in a controlled setting such as a usability laboratory or they can be undertaken 'in the wild' where much more realistic interactions will happen.
- **Data analytic**s can be gathered and analysed once a system or service is implemented.

2. Data analytics

Data analytics

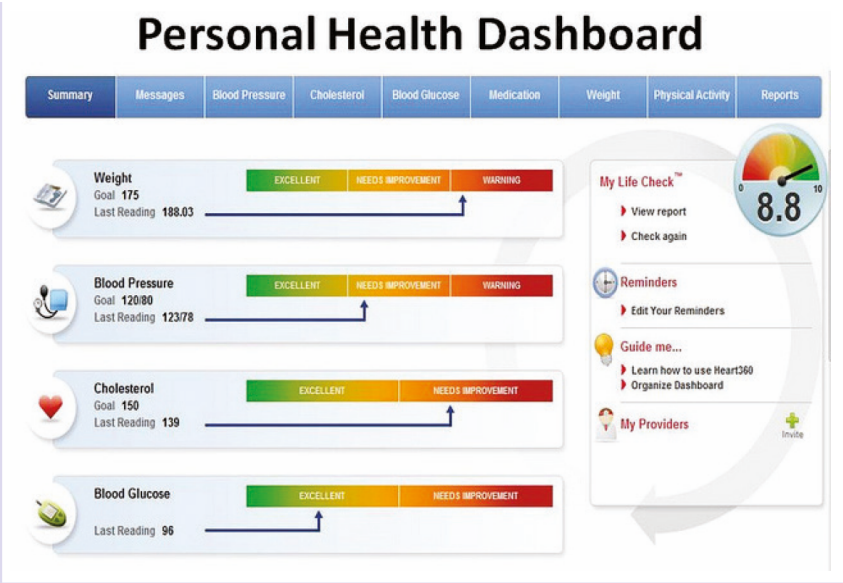
- Data analytics provides designers with data on system performance and the behaviours of individuals in interacting with systems and services.
- Data analytics also provides designers with interesting visualizations of the data and tools to help manipulate and analyse the data.
- The best known data analytics provider is **Google Analytics**, a free service that provides data about where users to websites and apps have come from (including their country, and potentially more detailed information about location and the device they were using) and what they did when they interacted with the system (such as how long they used the system, which pages of a site they visited, the order that they viewed pages and so on).

Data analytics

- We live in the era of '**big data**'. Huge amounts of data are being generated across many different fields. The **Internet of Things (IoT)** refers to the interconnectedness of sensors and devices with one another and across the internet. Through these connections, vast amounts of data are gathered and processed potentially providing new insights into many aspects of our environment.
- **Mobile devices** are collecting increasing amounts of personal data such as how many steps someone has taken in a day.
- Other **sensors** measure heart rate, blood pressure or levels of excitement in a person, this data has the potential to provide new insights into people's behaviours and performance.

Example: Personal analytics

- The availability of various **bio-sensors** in mobile and wearable devices has led to a movement known as **the quantified self or personal analytics**.
- Frequently associated with trying to get people to behave in a more healthy way, QS poses interesting questions about data gathering and use.
- For example, a watch will vibrate if a wearer has not stood up or moved around for an hour.
- It monitors and displays heart rate data.



- Other personal data such as the number of steps someone has taken in a day or the number of stairs they have climbed are presented on **personal 'dashboard'** visualizations.
- How people react to these various representations of themselves is an interesting issue (e.g. see Choe *et al*., 2014).

Analytics for Apps dashboard

Facebook-Google analytics

- **Facebook analytics** for apps is a free service that can be installed and provides information about who used an app on Facebook.
- Since users on Facebook have often provided a lot of personal information, more details of the users can be found.
- **Google Analytics** can provide demographic information based on what users have told them, sing a similar formula as that used to target Google Ads (advertisements).
- The data from Google or Facebook analytics is displayed using a ‘dashboard’.

Use of web analytics for evaluation

- Using data analytics services, designers can examine the activities of individuals and different groups such as Android phone users, people who accessed from a desktop machine using a particular browser and people who access the site from a particular location.
- Other important data for web analytics includes the number of visitors to a site over a period of time, the ‘bounce rate’ (the number of people who visited a site and then immediately left the site, without looking at any content), the number of pages viewed per session, time spent viewing pages and so on.

Other data analytics

- Other data analytic tools will provide a ‘heat map’ or a 'click map' of a website showing **which parts of a page are clicked on** most frequently.
- Other tools will allow the analyst to follow people’s browsing behaviour in real time, watching what they click on, how long they spend on particular sections.

Understanding users through data analytics

- The ability to understand user behaviour through data analytics, combined with the ability to rapidly deploy new versions of software, is changing the nature of interactive software development, as **developers can watch users behavior in real time**.
- In other circumstances, a company may issue its software with **two alternative interfaces** or with slightly different interfaces. The two interfaces are randomly assigned to users as they log onto a site. By looking at the analytics of the two interfaces, analysts can see which is performing better. This is known as **A/B testing** and is increasingly used to refine the UX of commercial websites.

https://www.smashingmagazine.com/2010/06/the-ultimate-guide-to-a-b-testing/

A/B testing example

Finding: Putting human photos on a website increases conversion rates by as much as double

https://www.smashingmagazine.com/2010/06/the-ultimate-guide-to-a-b-testing/

3. Expert evaluation

Expert evaluation

- A simple, relatively quick and effective method of evaluation is to get an **UX or usability expert** to inspect the service or system and try using it.
- Expert evaluation is no substitute for asking real people to use a design but expert evaluation is effective, particularly early in the design process.
- Experts will identify common problems based on their experience and factors that might otherwise interfere with an evaluation by non-experts.
- Although the methods have been around for over 20 years, expert based methods are still widely used by industry (Rohrer, Wendt, Sauro, Boyle, Cole, 2016).

Usability inspection methods

- Sometimes called usability inspection methods, there are a variety of approaches to expert evaluation
- An expert can simply be asked to look at a design and make suggestions.
- However, to help the experts structure their evaluation, it is useful to adopt a particular approach.
- This will help focus the expert's critique on the most relevant aspects for the purpose.
- The general approach to expert evaluation is that the expert will walk through **representative tasks or scenarios of use**.
- Additionally, they may adopt one of the **personas**. Thus, expert evaluation is tied to **scenario-based design** (and central to it).

Cognitive walkthrough

- Cognitive walkthrough is a rigorous paper-based technique for checking through the detailed design and logic of steps in an interaction.
- In essence, the cognitive walkthrough entails a usability or UX analyst stepping through the cognitive tasks that must be carried out in interacting with technology.
- Originally developed by Lewis *et al.* (1990) for applications where people browse and explore information, it has been extended to interactive systems in general (Wharton *et al.*, 1994).
- Aside from its systematic approach, the great strength of the cognitive walkthrough is that it is based on well-established theory rather than the trial and error or a heuristically based approach.

Cognitive walkthrough

- Inputs to the process are:
  - An understanding of the people who are expected to use the system.
  - A set of concrete **scenarios** representing both (a) very common and (b) uncommon but critical sequences of activities.
  - A complete description of the interface to the system
- The analyst asks the following four questions for each individual step in the interaction:
  - Will the people using the system **try to achieve the right effect**?
  - Will they notice that the correct action is **available**?
  - Will they **associate the correct action** with the effect that they are trying to achieve?
  - If the correct action is performed, will people see that **progress is being made** towards the goal of their activity?

Cognitive walkthrough

- If any of the questions is answered in the negative, then a usability problem has been identified and is recorded, but redesign suggestions are not made at this point.
- The process is carried out as a group exercise by **analysts and designers together**.
- The analysts step through usage scenarios and the design team are required to explain how the user would identify, carry out and monitor the correct sequence of actions (similar to program code walkthroughs).

Cognitive jogthrough

- The ‘cognitive jogthrough’ (Rowley and Rhoades, 1992) – **video records** (rather than conventional minutes) are made of walkthrough meetings
- They are annotated to indicate significant items of interest
- Design suggestions are permitted and low-level actions are aggregated wherever possible.

Streamlined Cognitive walkthrough

- The ‘streamlined cognitive walkthrough’ (Spencer, 2000) – the problem-free steps are not documented
- The four original questions are combined into two :
  - Will people **know what to do** at each step?
  - If people do the right thing, will they know that they did the right thing and **are making progress** towards their goal?
- Finally, the cognitive walkthrough is very often practised as a technique executed **by the analyst alone**, to be followed in some cases by a meeting with the design team.

Usability evaluation

- Most of the expert-based evaluation methods focus on the usability of systems.
- There are heuristics specific for websites or particular types of websites such as e-commerce sites.
- However, there is no problem with designers devising their own heuristics that focus on particular aspects of the UX that they are interested in.

Heuristic evaluation

- Heuristic evaluation refers to a number of methods in which a person trained in HCI, examines a proposed design to see how it measures up against a list of principles, guidelines or '**heuristics**' for good design.
- There are many **sets of heuristics** to choose from, both general-purpose and those relating to particular application domains, for example heuristics for web design.
- Ideally, **several people** with expertise in interactive systems design should review the interface.
- Each expert notes the problems and the relevant heuristic and **suggests a solution** where possible.

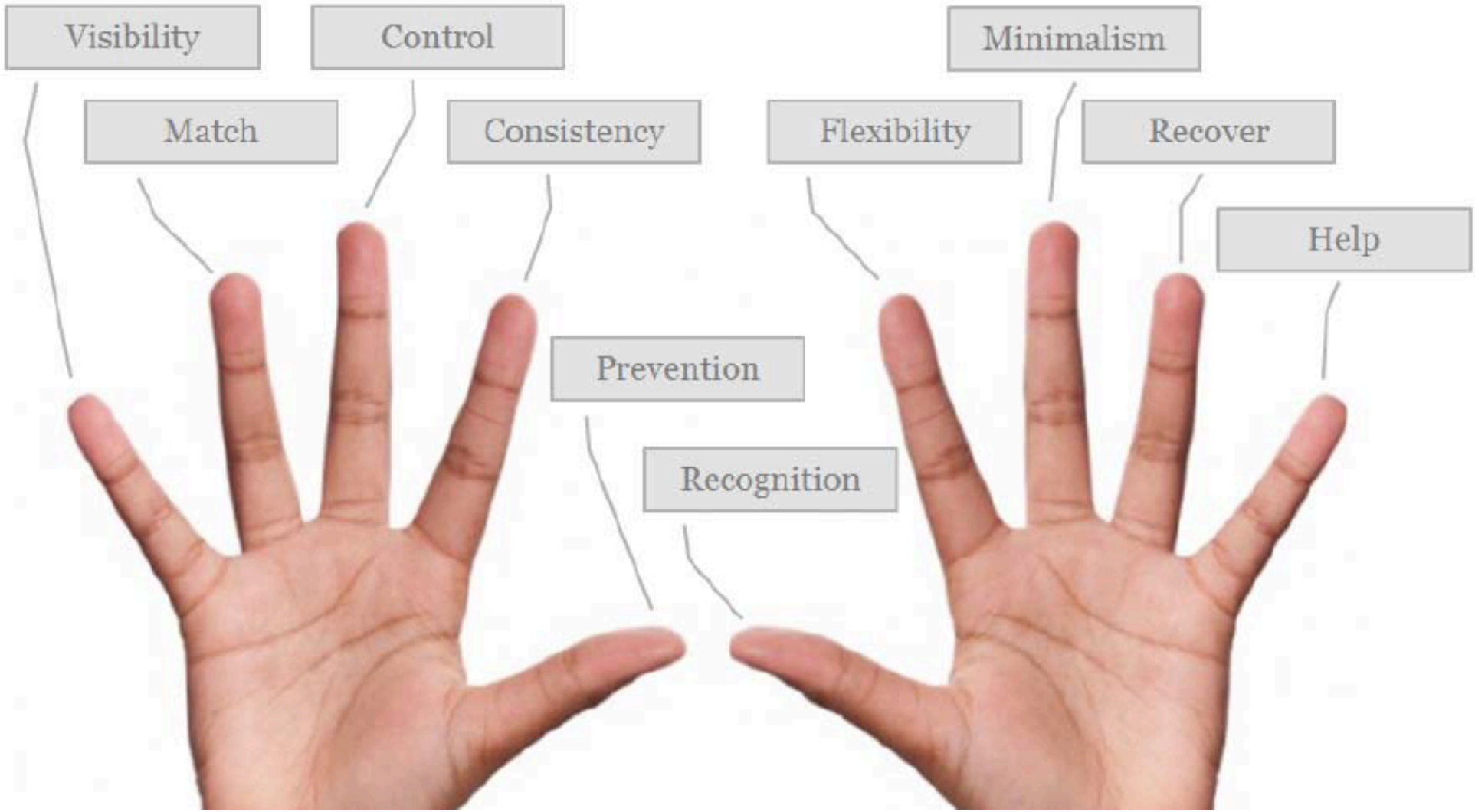
Heuristic evaluation

- It is helpful to add a **severity rating**, e.g. on a scale of 1 to 3, according to the likely impact of the problem, as recommended by Dumas and Fox (2012) in their comprehensive review of usability testing.
- However, they also note the disappointing level of correlation amongst experts in rating severity of problems.
- Evaluators work independently and then combine results.
- They may need to work through any training materials and be briefed by the design team about the functionality.

Heuristic evaluation: formative or summative evaluation?

- Heuristic evaluation is valuable as **formative evaluation** to help the designer improve the interaction at an early stage, and should not be used as **summative assessment** to make claims about the quality of a finished product.
- If that is what we need to do, then we must carry out properly designed and **controlled experiments involving greater number of participants**. However, the more controlled the testing situation becomes, the less it is likely to resemble the real world, which leads us to the question of '*ecological validity*'.

Heuristic Evaluation (Nielsen 1994)



Heuristic Evaluation (Nielsen 1994)

#1: Visibility of system status The system should always keep users informed, through appropriate feedback within reasonable time. (video)

#2: Match between system and the real world The system should speak the users' language, with words, phrases and concepts familiar to the user, (video)

#3: User control and freedom Users often need a clearly marked "emergency exit". Support undo and redo.(video.)

#4: Consistency and standards Users should not have to wonder whether different words, or actions mean the same thing. (video.)

#5: Error prevention Even better than good error messages is a careful design which prevents a problem from occurring in the first place. (video.)

#6: Recognition rather than recall Minimize the user's memory load by making objects, actions, and options visible. (video.)

#7: Flexibility and efficiency of use Accelerators — unseen by the novice user — may often speed up the interaction for the expert user (video.)

#8: Aesthetic and minimalist design Dialogues should not contain information which is irrelevant or rarely needed. (video.)

#9: Help users recognize, diagnose, and recover from errors Error messages should be expressed in plain language (no codes), precisely indicate the problem, and constructively suggest a solution. (video.)

#10: Help and documentation should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large. (video)

https://www.nngroup.com/articles/ten-usability-heuristics/

#1: Visibility
#2: Match
#3: Control
#4: Consistency
#5: Prevention
#6: Recognition
#7: Flexibility
#8: Minimalism
#9: Recover
#10: Help

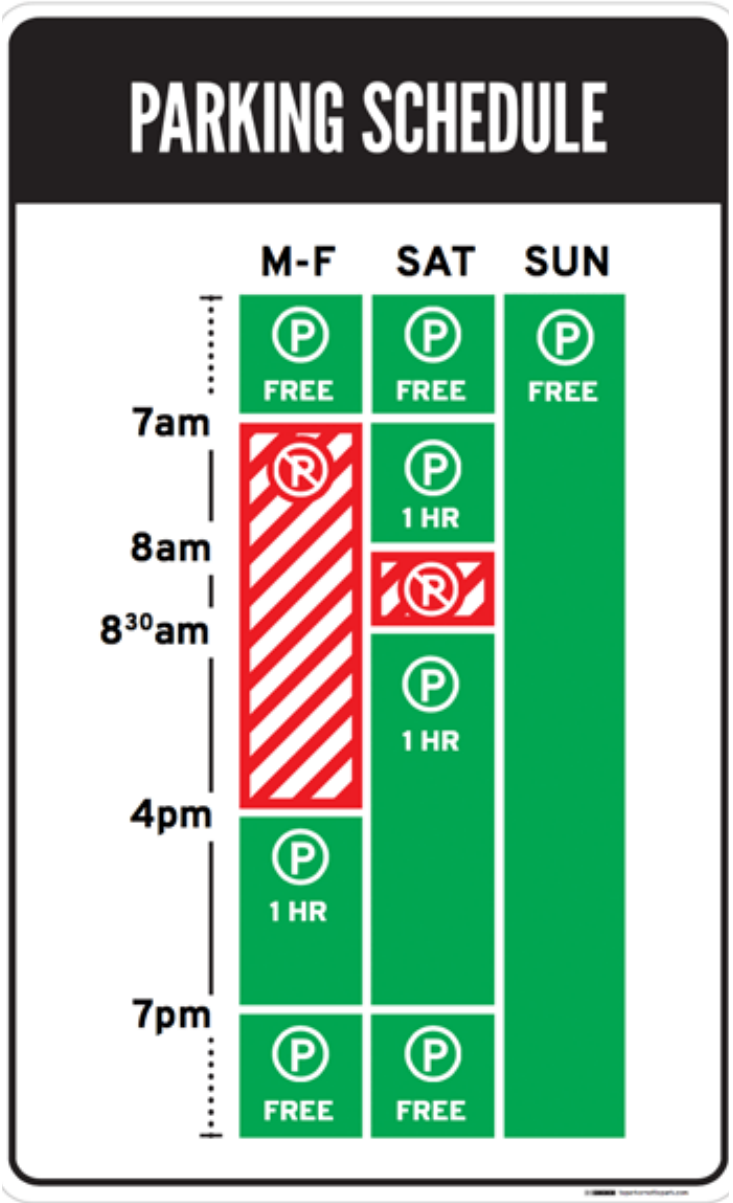
Severity rating

0 — don't agree it is a usability problem
1 — Cosmetic problem
2 — Minor usability problem
3 — Major usability problem; important to fix
4 — Usability catastrophe; imperative to fix

https://youtu.be/6Bw0n6Jvwxk

Which heuristic is violated? – redesign!

#1: Visibility
#2: Match
#3: Control
#4: Consistency
#5: Prevention
#6: Recognition
#7: Flexibility
#8: Minimalism
#9: Recover
#10: Help



https://www.interaction-design.org/literature/article/bad-design-vs-good-design-5-examples-we-can-learn-frombad-design-vs-good-design-5-examples-we-can-learn-from-130706

heuristic violated?

PayPal email receipt, posted on Dribbble

#1: Visibility
#2: Match
#3: Control
#4: Consistency
#5: Prevention
#6: Recognition
#7: Flexibility
#8: Minimalism
#9: Recover
#10: Help

1.Animate deliberately: think through each animation before you create it.
2.It takes more than 12 principles: Disney's 12 principles of animation work for films, but not necessarily for websites and apps.
3.Useful and necessary, *then* beautiful easthetics should of second priority.
4.Go four times faster: good animations are unobtrusive, they run fast.
5.Install a kill switch: for large animations, create an opt-out button.

https://www.interaction-design.org/literature/article/bad-design-vs-good-design-5-examples-we-can-learn-frombad-design-vs-good-design-5-examples-we-can-learn-from-130706

heuristic evaluation vs user testing

- Woolrych and Cockton (2000) conducted a large-scale **trial of heuristic evaluation**. Evaluators were trained to use the technique, then evaluated the interface to a drawing editor.
- The editor was then trialed by customers.
- Comparison of findings showed that many of the issues identified by the experts were not experienced by people (false positives), while some severe difficulties were missed by the inspection against heuristics. There were a number of reasons for this.

heuristic evaluation vs user testing

- Many false positives stemmed from a tendency by the experts to assume that people had no intelligence or even common sense.
- As for 'missing' problems, these tended to result from a series of mistakes and misconceptions often relating to a *set of linked items*, rather than isolated misunderstandings.
- Sometimes heuristics were misapplied or apparently added as an afterthought.
- Woolrych and Cockton conclude that the heuristics add little advantage to an expert evaluation and the results of applying them may be counter-productive. They (and other authors) suggest that more theoretically informed techniques such as the cognitive walkthrough offer more robust support for problem identification.
- It is very evident that heuristic evaluation is not a complete solution.
- At the very least, the technique must be used together with careful consideration of people and their real-life skills.
- **Participant evaluation** is required to get a realistic picture of the success of a system.

4. Participant-based evaluation

Participant-based evaluation

- Participant evaluation involves real people (users) in the evaluation.
- The participant methods range from designers sitting with participants as they work through a system to leaving people alone with the technology and observing what they do through a two-way mirror.

Real life vs Controlled studies

- REAL LIFE: People switch channels and interleave activities. They multitask, use several applications in parallel or in quick succession, are interrupted, improvise, ask other people for help, use applications intermittently and adapt technologies for purposes the designers never imagined.

Real life vs Controlled studies

- EVALUATION LIFE: The focus of most evaluations are small tasks usually part of lengthy sequences which change according to circumstances. Sequencies are extremely difficult to reproduce in testing and is often deliberately excluded from expert evaluations. So, the results of most evaluation studies is only indicative of issues in real-life usage.

Ecological validity

- Ecological validity is concerned with making an evaluation as life-like as possible.
- Designers can create circumstances that are as close to the real life environment as possible when undertaking an evaluation.
- Designs that appear robust in controlled, ‘laboratory’ settings can perform much less well in real-life, stressed situations.

Cooperative evaluation

- Andrew Monk and colleagues (Monk *et al.*, 1993) at the University of York (UK) developed cooperative evaluation as a means of maximizing the data gathered from a simple testing session.
- The technique is ‘cooperative’ because participants are not passive subjects but work as co-evaluators.
- It has proved a reliable but economical technique in diverse applications.

Guidelines for cooperative evaluation

Step	Notes
1 Using the scenarios prepared earlier, write a draft list of tasks.	Tasks must be realistic, doable with the software, and explore the system thoroughly.
2 Try out the tasks and estimate how long they will take a participant to complete.	Allow 50 per cent longer than the total task time for each test session.
3 Prepare a task sheet for the participants.	Be specific and explain the tasks so that anyone can understand.
4 Get ready for the test session.	Have the prototype ready in a suitable environment with a list of prompt questions, notebook and pens ready. A video or audio recorder would be useful here.
5 Tell the participants that it is the system that is under test, not them; explain and introduce the tasks.	Participants should work individually – you will not be able to monitor more than one participant at once. Start recording if equipment is available.
6 Participants start the tasks. Have them give you running commentary on what they are doing, why they are doing it and difficulties or uncertainties they encounter.	Take notes of where participants find problems or do something unexpected, and their comments. Do this even if you are recording the session. You may need to help if participants are stuck or have them move to the next task.
7 Encourage participants to keep talking.	Some useful prompt questions are provided below.
8 When the participants have finished, interview them briefly about the usability of the prototype and the session itself. Thank them.	Some useful questions are provided below. If you have a large number of participants, a simple questionnaire may be helpful.
9 Write up your notes as soon as possible and incorporate into a usability report.	

From Appendix 1 in Monk *et al.* (1993).

Sample questions during cooperative evaluation

Sample questions *during* the evaluation:

- What do you want to do?
- What were you expecting to happen?
- What is the system telling you?
- Why has the system done that?
- What are you doing now?

Sample questions *after* the session:

- What was the best/worst thing about the prototype?
- What most needs changing?
- How easy were the tasks?
- How realistic were the tasks?
- Did giving a commentary distract you?

From Appendix 1 in Monk *et al.* (1993).

Participatory heuristic evaluation

- The developers of participatory heuristic evaluation (Muller *et al.*, 1998) claim that it extends the power of heuristic evaluation without adding greatly to the effort required.
- An expanded list of heuristics is provided, based on those of Nielsen and Mack (1994) – One could use any heuristics
- The procedure for the use of participatory heuristic evaluation is just as for the expert version, but the participants are involved as ‘work-domain experts’ alongside usability experts and must be briefed about what is required.

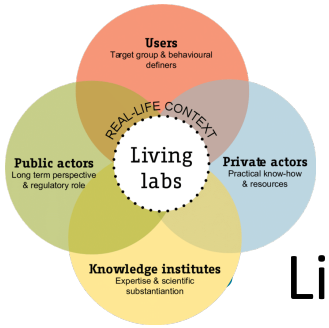
Co-discovery

- A naturalistic, informal technique, good for capturing first impressions. It is best used in the later stages of design.
- The standard approach of watching individual people interacting with the technology, and possibly ‘thinking aloud’ as they do so, can be varied by having participants **explore new technology in pairs**.
- For example, a series of pairs of people could be given a prototype of a new digital camera and asked to experiment with its features by taking pictures of each other and objects in the room.
- This tends to elicit a more naturalistic flow of comment, and people will often encourage each other to try interactions that they might not have thought of in isolation.

Co-discovery

- Depending on the data to be collected, the evaluator can take an active part in the session by asking questions or suggesting activities, or simply monitor the interaction either live or using a video recording.
- Inevitably, asking specific questions skews the output towards the evaluator's interests, but does help ensure that all important angles are covered.

- The term 'co-discovery' originates from Kemp and van Gelderen (1996) who provide a detailed description of its use.



Living Labs

Living Labs is a European approach to evaluation that aims to **engage as many people as possible** in exploring new technologies.

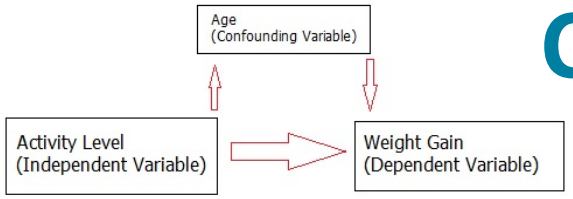
- For example, mobile phone manufacturers have teamed up with academics to hand out hundreds of early prototype systems to students to see how they use them.
- Other labs work with **elderly people** in their homes to explore new types of home technologies.
- Others work with **travellers and migrant workers** to uncover what new technologies can do for them.
- The key idea behind Living Labs is that people are both willing and able to contribute to designing new technologies and new services and it makes sense for companies to work with them.

Controlled experiments

- Another way of undertaking participant evaluation is to set up a controlled experiment.
- Controlled experiments are appropriate where the designer is interested in **particular features of a design**, perhaps comparing one design to another to see which is better.
- In order to do this with any certainty, the experiment needs to be carefully designed and run.
- The first thing to do when considering a controlled experiment approach to evaluation is to establish what it is that we are looking at (independent variable).
- For example, one might want to compare two different designs of a website, or two different ways of selecting a function on a mobile phone application.

Controlled experiments

- Once we have established what it is we are looking at, we need to decide how to measure the difference.
- These are the dependent variables.
- We might want to judge which web design is better based on the number of clicks needed to achieve some task; speed of access could be the dependent variable for selecting a function.
- Once the independent and dependent variables have been agreed, the experiment needs to be designed to avoid anything influencing the relationship between independent and dependent variables.



Controlled experiments: confounding variables

- Variables that can create bias and influence independent and dependent variables called confounding variables. These are **learning effects**, the effects of **different tasks**, the effects of different **background knowledge**.
- One possible reason for confounding variable is that the participants in any experiment are not balanced across the conditions. To avoid this, participants are usually divided up across the conditions so that there are roughly the same number of people in each condition and there are roughly the same number of males and females, young and old, and experienced and not.

Controlled experiments

- The next stage is to decide whether each participant will participate in all conditions (the so-called **within-subject design**) or whether each participant will perform in only one condition (the so-called **between-subject design**).
- In deciding this, we have to be wary of introducing confounding variables. For example, consider the learning effects that happen if people perform a similar task on more than one system.
- They start off slowly but soon get good at things, so if time to complete a task is a measure, they inevitably get quicker the more they do it.
- This effect can be controlled by randomizing the sequence in which people perform in the different conditions.

Controlled experiments combined with other studies

- The participants of a controlled experiment, are often asked to participate in other user studies.
- For example, an experiment being set up to look at **more than one independent variable**, perhaps one being looked at between subjects and another being looked at within subjects.
- Or **interviewing** the participants afterwards or using **focus groups** afterwards to find out other things about the design.
- People can be videoed and perhaps **talk aloud** during the experiments (so long as this does not count as a confounding variable) and this data can also prove useful for the evaluation.

Analysis of data of controlled experiments

- A controlled experiment will often result in some quantitative data: the measures of the dependent values.
- This data can then be analysed using statistics, for example comparing the average time across two conditions or the average number of clicks.
- See techniques discussed in the Research Methods course.

Challenge

- You have just completed a small evaluation project for a tourist information ‘walk-up-and-use’ kiosk designed for an airport arrivals area. A heuristic evaluation by you (you were not involved with the design itself) and a technical author found 17 potential problems, of which seven were graded severe enough to require some redesign and the rest were fairly trivial.
- You then carried out some participant evaluation. You had very little time for this, testing with only three people. The test focused on the more severe problems found in the heuristic evaluation and the most important functionality (as identified in the requirements analysis). Your participants – again because of lack of time and budget – were recruited from another section of your own organization which is not directly involved in interactive systems design or build but the staff do use desktop PCs as part of their normal work. The testing took place in a quiet corner of the development office.

Challenge

- Participants in the user evaluation all found difficulty with three of the problematic design features flagged up by the heuristic evaluation. These problems were essentially concerned with knowing what information might be found in different sections of the application. Of the remaining four severe problems from heuristic evaluation, one person had difficulty with all of them, but the other two people did not. Two out of the three test users failed to complete a long transaction where they tried to find and book hotel rooms for a party of travellers staying for different periods of time.
- What, if anything, can you conclude from the evaluation? What are the limitations of the data?

5. Evaluation in practice

Steps in evaluation projects

- The main steps in undertaking a simple but effective evaluation project are:
  - Establish the aims of the evaluation, the intended participants in the evaluation, the context of use and the state of the technology; obtain or construct scenarios illustrating how the application will be used.
  - Select **evaluation methods**. These should be a combination of expert-based review methods and participant methods.
  - Carry out **expert review**.
  - **Plan participant testing**; use the results of the expert review to help focus this.
  - **Recruit people** and organize testing venue and equipment.
  - Carry out the **evaluation**.
  - **Analyse results**, document and **report** back to designers.

Use of evaluation methods in practice

- A survey of 103 experienced practitioners of human-centred design conducted in 2000 (Vredenburg *et al.*, 2002) indicates that around 40 per cent of those surveyed conducted ‘usability evaluation’, around 30 per cent used ‘informal expert review’ and around 15 per cent used ‘formal heuristic evaluation’.

Aims of the evaluation (example)

- Deciding the aim(s) for evaluation helps determine the type of data required.
- Example: in the evaluation of a virtual training environment, the aims were to investigate the following:
  - Do the trainers understand and welcome the basic idea of the virtual training environment?
  - Would they use it to extend or replace existing training courses?
  - How close to reality should the virtual environment be?
  - What features are required to support record keeping and administration?

Aims of the evaluation

- If the aim of the evaluation is the **comparison of two different evaluation designs**, then much more focused questions will be required and the data gathered will be more quantitative. In the virtual training environment, for example, some questions we asked were:
  - Is it quicker to reach a particular room in the virtual environment using mouse, cursor keys or joystick?
  - Is it easier to open a virtual door by clicking on the handle or selecting the 'open' icon from a tools palette?
- Underlying issues were the focus on speed and ease of operation.
- With questions such as these, we are likely to need quantitative (numerical) data to support design choices.

Metrics and measures

Usability objective	Effectiveness measures	Efficiency measures	Satisfaction measures
Overall usability	• Percentage of tasks successfully completed • Percentage of users successfully completing tasks	• Time to complete a task • Time spent on non-productive actions	• Rating scale for satisfaction • Frequency of use if this is voluntary (after system is implemented)
Meets needs of trained or experienced users	• Percentage of advanced tasks completed • Percentage of relevant functions used	• Time taken to complete tasks relative to minimum realistic time	• Rating scale for satisfaction with advanced features
Meets needs for walk up and use	• Percentage of tasks completed successfully at first attempt	• Time taken on first attempt to complete task • Time spent on help functions	• Rate of voluntary use (after system is implemented)
Meets needs for infrequent or intermittent use	• Percentage of tasks completed successfully after a specified period of non-use	• Time spent re-learning functions • Number of persistent errors	• Frequency of reuse (after system is implemented)
Learnability	• Number of functions learned • Percentage of users who manage to learn to a pre-specified criterion	• Time spent on help functions • Time to learn to criterion	• Rating scale for ease of learning

Source: ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals (VDTs), extract of Table B.2

Metrics and measures

- In most of these, there is a task – something the participant needs to get done – and it is reasonably straightforward to decide whether the task has been achieved successfully or not.
- There is one major difficulty: deciding the acceptable figure for, say, the percentage of tasks successfully completed. Is this 95 per cent, 80 per cent or 50 per cent?
- Otherwise, a **baseline** may be available from comparative testing against an alternative design, a previous version, a rival product or the current manual version of a process to be computerized.
- But the evaluation team still has to determine whether **a metric is relevant**.
- For example, in a complex computer-aided design system, one would not expect most functions to be used perfectly at the first attempt.
- Speed of keying characters may be crucial to the success of a mobile phone.

Metrics and measures

- Factors to consider in deciding metrics:
  - Just because something can be measured, it doesn't mean it should be measured.
  - Always we should refer back to the overall purpose and context of use of the technology.
  - Consider the usefulness of the data we are likely to obtain against the resources it will take to test against the metrics.

Evaluating entertainment apps

- What metrics to use for evaluation of Games and other applications designed for entertainment ?.
- While we may still want to evaluate whether the basic functions (to move around a game environment, for example), are **easy to learn**, efficiency and effectiveness in a wider sense are much less relevant.
- The ‘purpose’ here is **to enjoy the game** and time to complete, for example, a particular level may sometimes be less important than **experiencing the events** that happen along the way.

Evaluating engagement

- Similarly, multimedia applications are often directed at intriguing users or evoking **emotional responses** rather than having the achievement of particular tasks in a limited period of time.
- In contexts of this type, evaluation centres on probing user experience through **interviews or questionnaires**.
- Read and MacFarlane (2000), for example, used a rating scale presented as a ‘smiley face vertical fun meter’ when working with children to evaluate novel interfaces.
- Other measures which can be considered are observational: the **user’s posture or facial expression**, for instance, may be an indicator of engagement in the experience.

Participants in evaluation

- The most important people in evaluation are the people who will use the system.
- Analysis work should have identified the characteristics of these people and represented these in the form of **personas**.
- In particular, we should know the **skills** relating to input and output devices, **experience, education, training and physical and cognitive capabilities**.
- Relevant data can include knowledge of the **activities** the technology is intended to support.

How many participants ?

- We need to recruit at least three and preferably five people to participate in tests.
- Nielsen's recommended sample of 3–5 participants has been accepted wisdom in usability practice for over a decade. However, some practitioners and researchers advise that this is too few.
- We consider that in many real-world situations, obtaining even 3–5 people is difficult, so small test numbers are recommended as part of a pragmatic evaluation strategy.
- If we have a heterogeneous set of customers that the design is aimed at, then we will need to run 3–5 people from each group through the user tests.

Recruiting participants

- Finding representative participants should be straightforward if we are developing an in-house application.
- Otherwise, participants can be found through focus groups established for marketing purposes or, if necessary, through advertising.
- Students are often readily available but they are only representative of a particular segment of the population.
- If we have adequate resources, payment can help recruitment. Inevitably, the sample will be biased towards cooperative people with some sort of interest in technology, so we should bear this in mind when interpreting the results.

Recruiting participants

- If we cannot recruit any genuine participants – representative of the target customers – at least we should have someone else try to use it.
- This could be one of our colleagues, a friend, a relative or anyone we trust to give us an honest reaction.
- Almost certainly, they will find some design flaws.
- The data we obtain will be limited but better than nothing.
- We should however, have to be extremely careful as to how far you generalize from your findings.

How much help during testing?

- Evaluators set up the tests and collect data but how far they should become involved?
- The recommended method for basic testing requires an evaluator to sit with each user and engage with them as they carry out the test tasks.
- It is suggested that for ethical reasons and in order to keep the tests running, evaluators should provide help if the participant is becoming uncomfortable or completely stuck.
- The amount of help that is appropriate will depend on the type of application (e.g. for an information kiosk for public use, we might provide only very minimal help), the degree of completeness of the test application and, in particular, whether any help facilities have been implemented.

Tracking eye movement

- Eye-movement tracking (‘eye tracking’) can show participants’ changing focus on different areas of the screen.
- This can indicate which features of a user interface have attracted attention, and in which order, or capture larger-scale gaze patterns indicating how people move around the screen.

- Eye tracking is very popular with website designers as it can be used to highlight which parts of the page are most looked at, the ‘hot spots’, and which are missed altogether.
- Eye-tracking software is readily available to provide maps of the screen.
- Some of it can also measure pupil dilation, which is taken as an indication of arousal.
- Physiological techniques in evaluation rely on the fact that all our emotions – anxiety, pleasure, apprehension, delight, surprise and so on – generate physiological changes.

Physical and physiological measures

- The most common measures are of changes in **heart rate**, the **rate of respiration**, **skin temperature**, **blood volume**, pulse and galvanic skin response (an indicator of the amount of perspiration).
- All are indicators of changes in the overall level of arousal, which in turn may be evidence of an **emotional reaction**.
- Sensors can be attached to the participant's body (commonly the fingertips) and linked to software which converts the results to numerical and graphical formats for analysis.
- There are many **unobtrusive methods** too, such as **pressure sensors** in the steering wheel of a games interface or sensors that measure if the participant is on the edge of his/her seat.

Measuring emotions

- Which particular emotion is being evoked cannot be deduced from the **level of arousal** alone but must be inferred from other data such as **facial expression, posture or direct questioning**.
- Typically, startling events or threatening features are produced in the environment and arousal levels measured as people encounter them.
- Researchers at University College London and the University of North Carolina at Chapel Hill (Usoh *et al.*, 1999, 2000; Insko, 2001, 2003; Meehan, 2001) have conducted a series of experiments when measuring arousal as participants approach a '**virtual precipice**'.

Virtual reality cliff

Source: Reprinted from *Being There: Concepts, Effects and Measurement of User Presence in Synthetic Environment,* Inkso, B.E., Measuring presence. © 2003, with permission from IOS Press

Physical and physiological measures

- Galvanic skin response (GSR), measures the level of arousal that a person is experiencing. A sensor placed on the user's skin will record how much perspiration there is and hence how aroused the person is.
- Face recognition can determine if people are looking happy or sad, confused or angry.
- The Facial Action Coding System (FACS) is a robust way of measuring emotion through facial expression.
- Pressure sensors can detect how tightly people are gripping something. Of course, video can be used to record what people are doing.
- These various measures can be combined into a powerful way of evaluating UX.

Measuring presence

- There techniques involving physiological measures for the assessment of the degree of presence – the sense of 'being there' evoked by virtual environments
- Presence (a shortened version of the term "telepresence") is a psychological state or subjective perception in which even though part or all of an individual's current experience is generated by and/or filtered through human-made technology, **part or all of the individual's perception fails to accurately acknowledge the role of the technology in the experience**

Evaluating presence

- Designers of virtual reality (VR) and augmented reality (AR) applications are often concerned with the sense of presence, of being ‘there’ in the virtual environment rather than ‘here’ in the room where the technology is being used.
- A strong sense of presence is thought to be crucial for such applications as games, those designed to treat phobias, to allow people to ‘visit’ real places they may never see otherwise or indeed for some workplace applications such as training to operate effectively under stress.
- The sense of presence is strongly entangled with individual dispositions, experiences and expectations. Of course, this is also the case with reactions to any interactive system but presence is an extreme example of this problem.
- The concept of presence itself is ill-defined and the subject of much debate amongst researchers. Variants include the sense that the virtual environment is realistic, the extent to which the user is impervious to the outside world, the retrospective sense of having visited rather than viewed a location and a number of others.
- Asking people about presence while they are experiencing the virtual environment tends to interfere with the experience itself. On the other hand, asking questions retrospectively inevitably fails to capture the experience as it is lived.

Evaluating presence: difficulties

- In one experiment, questionnaire results showed that while many people did not feel wholly present in the virtual environment (a recreation of an office), some of them did not feel wholly present in the real-world office either (Usoh *et al.*, 2000).
- Less structured attempts to capture verbal accounts of presence include having people write accounts of their experience or inviting them to provide free-form comments in an interview.
- The results are then analysed for indications of a sense of presence.
- The difficulty here lies in defining what should be treated as such an indicator, and in the layers of indirection introduced by the relative verbal dexterity of the participant and the interpretation imposed by the analyst.
- Other approaches to measuring presence attempt to avoid such layers of indirection by observing behaviour in the virtual environment or by direct physiological measures.

6. Evaluation plan and tools used

The test plan and task specification

- A plan should be drawn up to guide the evaluation. The plan specifies:
  - Aims of the test session
  - Practical details, including where and when it will be conducted, how long each session will last, the specification of equipment and materials for testing and data collection, and any technical support that may be necessary
  - Numbers and types of participant
  - Tasks to be performed, with a definition of successful completion. This section also specifies what data should be collected and how it will be analysed.
- You should now conduct a **pilot session** and fix any unforeseen difficulties. For example, task completion time is often much longer than expected and instructions may need clarification.

Reporting usability evaluation results to the design team

- The evaluation is worthwhile if the results are acted upon.
- An organized list of findings is needed in order to prioritize redesign work.
- If we are reporting back to a design/development team, it is crucial that they we can see immediately what the problem is, how significant its consequences are and ideally what needs to be done to fix it.
- The report should be ordered either by areas of the system concerned or by severity of problem.
- For the latter, we could adopt a three- or five-point scale, perhaps ranging from ‘would prevent participant from proceeding further’ to ‘minor irritation’.
- Adding a note of the general usability principle concerned may help designers to understand why there is a difficulty but often more specific explanation will be needed.

Reporting usability evaluation results to the design team

- A face-to-face meeting may have more impact than a written document alone (although this should always be produced as supporting material)
- This would be the ideal venue for showing short video clips of participant problems. Suggested solutions make it more probable that something will be done.
- If the organization has a formal quality system, an effective strategy is to have usability evaluation alongside other test procedures, so usability problems are dealt with in the same way as any other fault.
- Usability problems can be fed into a ‘bug’ reporting system if one exists.

Evaluating usability / SUS

- There are several standard ways of measuring usability but probably the best known and most robust is the system usability scale (SUS).
- Jeff Sauro presents the scale as. He suggests that any score over 68 is above average and indicates a reasonable level of usability.

The System Usability Scale

The SUS is a 10 item questionnaire with 5 response options.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The SUS uses the following response format:

Strongly Disagree 1	2	3	4	Strongly Agree 5
○	○	○	○	○

Scoring SUS

- For odd items: subtract one from the user response.
- For even-numbered items: subtract the user responses from 5
- This scales all values from 0 to 4 (with four being the most positive response).
- Add up the converted responses for each user and multiply that total by 2.5. This converts the range of possible values from 0 to 100 instead of from 0 to 40.

Evaluating UX

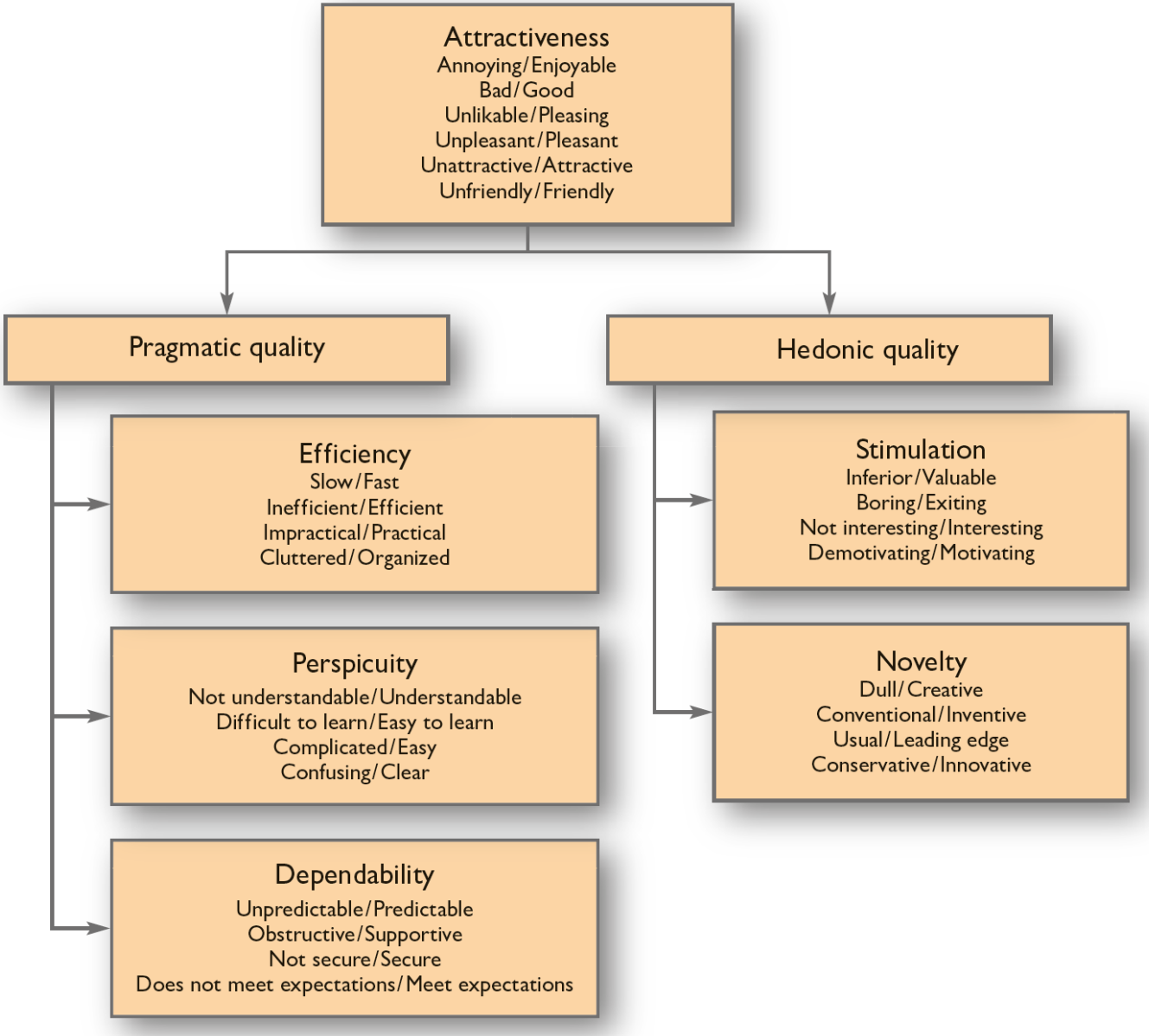
- There are a number of tools and methods specifically aimed at evaluating user experience.
- They differentiate between the pragmatic qualities of the UX and the hedonic qualities (Hassenzahl, 2010).
- The **user experience questionnaire** describes these qualities. It is a 26 item questionnaire, used to gather data about a UX
- An alternative is the **AttracDiff questionnaire**

User experience questionnaire

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

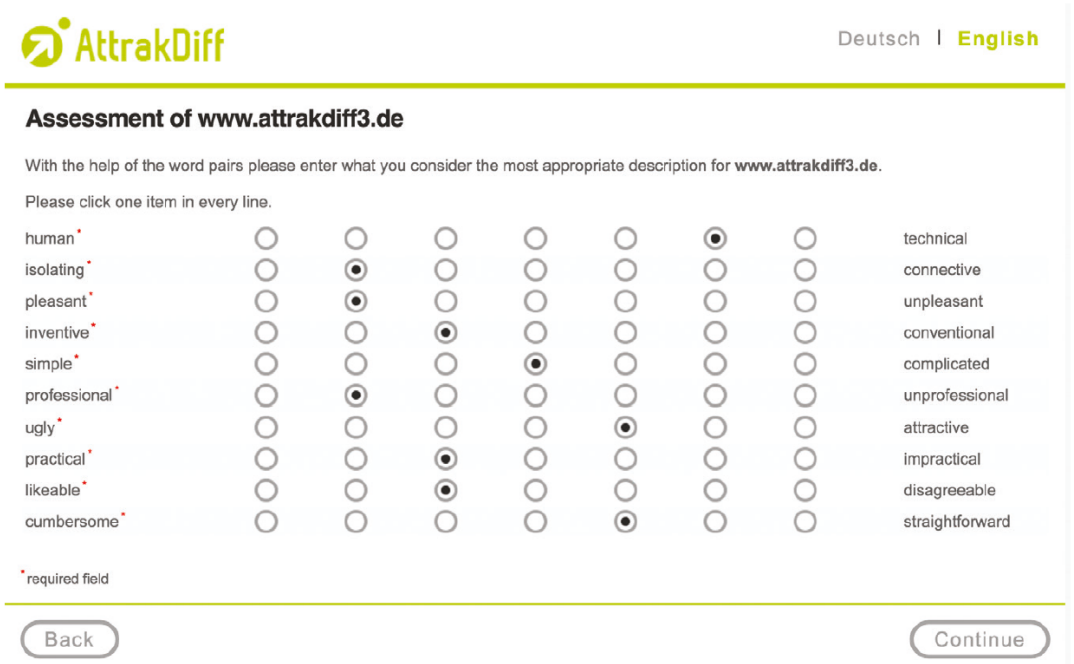
semantic differentials

User experience questionnaire



AttrakDiff: evaluating US

- An alternative to UEQ is to use the Attrakdiff on-line questionnaire.
- This has a similar approach but uses different terms.
- Both of these questionnaires can be used as they are and this has the advantage that comparisons can be made across products and services.
- For specific evaluation, however, UX designers may need to change the terms used on the semantic differential scales.



Tools for Evaluating presence

- The measures used in evaluating presence adopt various strategies to avoid these problems but none are wholly satisfactory.
- The various questionnaire measures, for example, the questionnaire developed by NASA scientists Witmer and Singer (1998) or the range of instruments developed at University College and Goldsmiths College, London (Slater, 1999; Lessiter *et al.*, 2001), can be cross-referenced to measures which attempt to quantify how far a person is generally susceptible to being 'wrapped up' in experiences mediated by books, films, games and so on as well as through virtual reality.
- The Sense of Presence Inventory (SOPI) can be used to measure media presence.
- The Witmer and Singer Immersive Tendencies Questionnaire (Witmer and Singer, 1998) is the best known of such instruments.
- However, presence as measured by presence questionnaires is a slippery and ill-defined concept.

Evaluation at home

- People at home pose new challenges for the evaluator compared to those at work. What do you think the challenges are?
- They are likely to be more concerned about **protecting their privacy** and generally unwilling to spend **their valuable leisure time** in helping you with your usability evaluation.
- So, it is important that data gathering techniques are interesting and stimulating for users and make as little demand on time and effort as possible.
- Petersen *et al.* (2002), for example, were interested in the evolution over time of relationships with technology in the home.
- They used conventional **interviews** at the time the technology (a new television) was first installed but followed this by having families act out scenarios using it.
- **Diaries** were also distributed as a data collection tool, but in this instance, the non-completion rate was high possibly because of the complexity of the diary pro forma and the incompatibility between a private diary and the social activity of television viewing.

Evaluation at home

- Where the family is the focus of interest, techniques should be engaging for children as well as adults – not only does this help to ensure that all viewpoints are covered but also working with children is a good way of drawing parents into evaluation activities.
- An effective example of this in early evaluation is reported in Baillie *et al.* (2003) and Baillie and Benyon (2008), in which the investigators supplied users with Post-its to capture their thoughts about design concepts. An illustration of each different concept was left in the home in a location where it might be used and users were encouraged to think about how they would use the device and any issues that might arise.

Summary

- This chapter has presented an overview of the key issues in evaluation.
- Designing the evaluation of an interactive system, product or service requires as much attention and effort as designing any other aspect of that system.
- Designers need to be aware of the possibilities and limitations of different approaches and, in addition to studying the theory, they need plenty of practical experience.
- Designers need to focus hard on what features of a system or product they want to evaluate.
- They need to think hard about the state that the system or product is in and hence whether they can evaluate those features.

Key points

- Designers can gather and study data analytics on the performance of their service.
- We have reviewed expert-based methods of evaluation.
- We also looked at participant-based methods of evaluation.
- Designers need to design their evaluation to fit the particular needs of the contexts of use and the activities that people are engaged in.